Comparing chemical fingerprints for ecotoxicology

Leander Schietgat[1], Bertrand Cuissart[2], Alban Lepailleur[3], Kurt De Grave[1], Bruno Crémilleux[2], Ronan Bureau[3], and Jan Ramon[1]

[1] Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium
[2] Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen, UMR CNRS 6072, Université de Caen Basse-Normandie
[3] Centre d'Études et de Recherche sur le Médicament de Normandie (CERMN), EA 4258-INC3M FR CNRS 3038, Université de Caen Basse-Normandie

Abstract. We present a comparison between 14 chemical fingerprints using 1D/2D features and PMCSFG, a data mining method that induces features based on maximum common subgraphs. We provide an experimental evaluation and discuss the usefulness of the different methods on ecotoxicology data. The features generated by data mining yield a similar performance for predicting toxicity, while they are more interpretable by chemists.

1 Introduction

The evolution of data mining techniques provides methods for efficiently finding relations between chemicals and toxicological endpoints in large datasets. A fundamental part of any data mining study is how to encode the information. Molecular fingerprints [1], where structural features are represented by a Boolean array, are a standard and computationally efficient representation of chemical compounds. They have been successfully used in molecular similarity search [2, 3], and thus provide a relevant description of molecular structures. We focus on the use of 2D fingerprints in predictive toxicology and we compare different algorithms to construct them. These algorithms can be divided into four classes: (1) dictionary-based, (2) path-based, (3) radial-based, and (4) atom pair-based. As an alternative to the above approaches, we also considered a maximum common substructure (MCS) algorithm to describe the molecules. We trained predictive toxicology models from each of the chemical descriptions with several machine learning algorithms widely used in chemoinformatics [4]. The results suggest that the MCS-based chemical descriptors are an interesting alternative to the standard chemical fingerprints.

2 Fingerprint methods

The features generated by fingerprint methods are used to encode each molecule in a dataset as a k-dimensional binary vector (with k the number of features), where a 1 is marked in the i-th position if the i-th feature occurs in the example. We divide fingerprint methods into five categories, based on the way features are generated.

Dictionary-based fingerprints rely on features which have been identified *a priori* by domain experts as important fragments: MACCS keys consist of 166 features mostly encoded by SMARTS patterns, PubChem keys have 883 features corresponding to PubChem substructures. The remaining fingerprint types conceptually encode fragments based on the atom-bond structures in the dataset. **Path-based** fingerprints can enumerate all linear paths up to seven bonds including the description of rings up to 14 bonds (Linear), or the linear paths augmented with intersections of linear paths, with a maximum of five bonds per path to encode branched features (Dendritic), or features of four consecutively bonded non-hydrogen atoms along with the number of non-hydrogen branches, corresponding to a torsion angle (Torsion). **Radial-based** fingerprints iteratively encode features that represent each heavy atom in larger and larger structural neighborhoods, up to a given diameter (2, 4, and 6 in this study). The atom type (ECFP) and the functional class (FCFP) encoding rules were used to define the atom abstraction. In another approach, each heavy atom in a structure is characterized by an environment that consists of all other heavy atoms within a distance of two bonds. Each member of the list is encoded into a string of the form Type-freq(Type)-d, where freq(Type) is the number of times a given atom type is found at a distance d from the central atom. The atom-typing scheme used is the Sybyl Mol2 (MOLPRINT2D). **Atom pair-based** fingerprints encode features representing two atoms and their corresponding distance, for example the Carthart atom types and the topological distance separating them (Pairwise) or, as an extension, a triplet consisting of a set of three atoms and the topological distances separating them. As there are six different ways to order the atoms in a triplet, a canonicalization is performed, ensuring that every bit corresponds to a unique triplet (Triplets). **Pairwise Maximum Common Subgraph** Feature Generation (PMCSFG) [5] computes maximum common subgraphs (MCSs) under the block-and-bridge-preserving subgraph isomorphism between molecule pairs. For efficiency reasons the algorithm computes MCSs only from outerplanar graphs. For a given set of graphs, the method can compute all possible MCSs or a random subset [6].

3 Experiments

Dataset The *European Chemicals Bureau* (ECB) dataset was constructed using data from the European Chemicals Agency. We only kept chemicals annotated with standardized phrases implemented by the EU through the CLP regulation and referring to the hazard of the substance to aquatic organisms. The dataset was cleaned following standard practices such as the removal of inconsistent compounds or the addition of hydrogens on hetero-atoms. This resulted in a dataset with 372 chemicals annotated "very toxic" and 195 chemicals annotated "harmful".

Experimental methodology In order to compare the different fingerprint methods, we used them as features in five different machine learning methods: decision tree learning (DT), instance-based learning (IBL), naïve Bayes (NB), rule-based learning (RBL) and support vector machines (SVM). The learning task is to discriminate the harmful molecules from the toxic ones. We based the SVMs on the Tanimoto kernel [6], which computes a similarity between two fingerprints as the number of common features (i.e., the set intersection between the two molecules) divided by the total number of patterns that occur in either or both of the molecules (i.e., the set union). Tanimoto is the recommended kernel for fingerprints of small molecules. As implementation we used SVM^{light} and C-parameter tuning. For the other learning methods, we used the Weka data mining tool, with standard configurations. We report the performance as the area under the ROC curve.

Fingerprint method	# features	DT	IBL	NB	RBL	SVM	Average
MACCS	166	0.75	0.77	0.72	0.69	0.89	**0.76**
PubChem	883	0.76	0.79	0.75	0.65	0.90	**0.77**
Linear	42757	0.71	0.60	0.78	0.57	0.88	**0.71**
Dendritic	24933	0.70	0.66	0.77	0.57	0.85	**0.71**
Torsion	2051	0.63	0.58	0.68	0.58	0.74	**0.62**
ECFP$_2$	1111	0.73	0.76	0.78	0.58	0.88	**0.75**
ECFP$_4$	3821	0.71	0.70	0.79	0.59	0.87	**0.73**
ECFP$_6$	7058	0.71	0.70	0.79	0.55	0.87	**0.72**
FCFP$_2$	258	0.78	0.77	0.73	0.62	0.87	**0.75**
FCFP$_4$	1807	0.75	0.73	0.76	0.63	0.86	**0.75**
FCFP$_6$	4461	0.76	0.69	0.77	0.65	0.89	**0.75**
MOLPRINT2D	2163	0.74	0.72	0.80	0.58	0.88	**0.74**
AtomPairs	3778	0.73	0.69	0.75	0.58	0.86	**0.72**
Triplets	102447	0.67	0.68	0.75	0.56	0.84	**0.70**
PMCSFG	1218	0.77	0.76	0.73	0.62	0.89	**0.76**
Average	**13259**	**0.73**	**0.71**	**0.76**	**0.60**	**0.87**	

Table 1. Area under the ROC for the different fingerprint methods and learning algorithms on the ECB dataset.

Results Table 1 shows the results of the comparison. The Torsion fingerprints perform the worst on average, while PubChem keys, MACCS keys, and PMCSFG perform the best. Interestingly, the three best performing fingerprints have the fewest features (133 to 1218). The average performance of the different learning methods varies from 0.60 to 0.87. Quantitatively, *whatever the molecular description is*, SVM appears to be the best performing learning method while rule-based learning leads to the poorest results. Moreover, there is a significant difference between the average score of the SVMs and the second best learner, NB.

4 Discussion

The preliminary results show that the dictionary-based fingerprints (PubChem keys and MACCS keys) obtain the best performance on average, with MCS fingerprints following closely behind. Apart from an adequate predictive performance, MCS fingerprints have multiple advantages. First, they can be automatically and efficiently learned from data, unlike the dictionary-based fingerprints which are selected by hand. Second, the feature set is smaller than the non-dictionary-based fingerprints. Moreover, the MCS features are typically much larger than the ones in the former fingerprints (9 atoms on average, up to 28 atoms for the ECB dataset) and hence carry more information. Third, an MCS feature corresponds to a recognizable molecular fragment, which makes it directly significant for a chemist. During the talk, we will illustrate this by exhibiting some examples of the recovery of known structural alerts for ecotoxicity, e.g., organophosphorus moieties, chlorobenzenes or phenol rings.

Acknowledgments

This work has been achieved thanks to joint financial supports: Région Basse-Normandie, ERC Starting Grant 240186 "MiGraNT", Research Fund KU Leuven, IWT (SBO Nemoa, SBO InSPECtor).

References

1. Todeschini, R., Consonni, V.: Molecular Descriptors for Chemoinformatics. Wiley-VCH (2009)
2. Willett, P.: Similarity-based virtual screening using 2D fingerprints. Drug Discovery Today **11**(23-24) (2006) 1046–1053
3. Geppert, H., Bajorath, J.: Advances in 2D fingerprint similarity searching. Expert Opinion on Drug Discovery **5**(6) (2010) 529–542
4. Varnek, A., Baskin, I.: Machine learning methods for property prediction in chemoinformatics: Quo vadis? Journal of Chemical Information and Modeling **52**(6) (2012) 1413–1437
5. Schietgat, L., Ramon, J., Bruynooghe, M.: A polynomial-time maximum common subgraph algorithm for outerplanar graphs and its application to chemoinformatics. Ann. Math. Artif. Intel. (2013) 1–34
6. Schietgat, L., Costa, F., Ramon, J., De Raedt, L.: Effective feature construction by maximum common subgraph sampling. Machine Learning **83** (2011) 137–161